TECHNICAL RESEARCH REPORT

$M|G|\infty$ Input Process: A Versatile Class of Models for Network Traffic

by M. Parulekar, A.M. Makowski

T.R. 96-59

ISR
INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
1996

2. REPORT TYPE

3. DATES COVERED
00-00-1996 to 00-00-1996

4. TITLE AND SUBTITLE
M/G/ input processes: A versatile class of models for network traffic

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Electrical Engineering,Institute for Systems Research,University of Maryland,College Park,MD,20742

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
unclassified	**unclassified**	**unclassified**		**22**	

$M|G|\infty$ input processes: A versatile class of models for network traffic

MINOTHI PARULEKAR [†]
minothi@eng.umd.edu
(301) 405-2948

ARMAND M. MAKOWSKI [‡]
armand@eng.umd.edu
(301) 405-6648
FAX:(301) 314-9281

Abstract

We suggest the $M|G|\infty$ input process as a viable model for network traffic due to its versatility and tractability. To gauge its performance, we study the large buffer asymptotics of a multiplexer driven by an $M|G|\infty$ input process. We identify the process as short or long–range dependent by means of simple tests. The decay rate of the tail probabilities for the buffer content (in steady–state) at the multiplexer is investigated using large deviation techniques suggested by Duffield and O'Connell. The appropriate large deviations scaling is found to be related to the forward recurrence time for the service time distribution, and a closed–form expression is derived for the corresponding generalized limiting log–moment generating function associated with the input process. Two very different regimes are identified. We apply our results to cases where the service time distribution in the $M|G|\infty$ input model is (i) Rayleigh (ii) Gamma (iii) Geometric (iv) Weibull (v) Log–normal and (vi) Pareto – cases (v) and (vi) have recently been found adequate for modeling packet traffic streams in certain networking applications. Finally, we comment

[†] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742. The work of this author was supported through NSF Grant NSFD CDR-88-03012 .

[‡] Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742. The work of this author was supported partially through NSF Grant NSFD CDR-88-03012, NASA Grant NAGW277S and the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0002.

on the insufficiency of the short or long–range dependence in the process in clearly describing buffer dynamics.

1 Introduction

Several recent measurement studies have concluded that classical Poisson–like traffic models are ill-equipped to account for time dependencies observed at multiple time scales in a wide range of networking applications, including Ethernet LANs [12, 16, 26], VBR traffic [13], WAN traffic [25]. As the resulting temporal correlations are expected to have a significant impact on buffer engineering practices, this "failure of Poisson modeling" has generated an increased interest in a number of alternative traffic models which capture observed (long–range) dependencies. Proposed models include the fractional Brownian motion input model [20] and the fractional Gaussian noise input process [1]; already both have exposed clearly the limitations of traditional traffic models in predicting storage requirements, congestion control and other measures of traffic performance.

In this paper we focus instead on the class of $M|G|\infty$ input processes as potential traffic models. An $M|G|\infty$ input process is understood as the busy server process of a discrete–time infinite server system fed by a discrete–time Poisson process of rate λ (customers/slot) and with generic service time σ distributed according to G. We argue that $M|G|\infty$ input processes provide a viable alternative to existing traffic models; reasons which range from flexibility to tractability, are briefly presented below. A lengthier discussion, given later in the body of the paper, is based mostly on the results developed in the papers [23, 24]; we refer the reader to these references for proofs as well as for additional information:

Firstly, the $M|G|\infty$ input model has been succesfully investigated as a model for some wide area applications, e.g., Paxson and Floyd [25] report a good fit to TELNET and FTP data using a log–normal service time [25]. However, the relevance of the $M|G|\infty$ input model to network traffic modelling is perhaps best explained through its connection to an attractive model for aggregate packet streams proposed by Likhanov, Tsybakov and Georganas [18]. They combine traffic generated by several on–off sources with a Pareto distributed activity period, and show that increasing the number of sources yields a limiting behaviour identical to the $M|G|\infty$ input stream with a Pareto distributed σ. As should be clear from their analysis, the limiting result holds for arbitrary activity period distributions, thereby providing a rationale for the view that $M|G|\infty$ input processes could provide a natural alternative to existing traffic models, at least for certain multiplexed applications. This limiting argument is similar to that of using the Palm–Khintchin Theorem to justify the Poisson model for interactive data traffic.

Secondly, the class of $M|G|\infty$ input processes has the desirable property of being stable under multiplexing, i.e., the superposition of several $M|G|\infty$ processes can be represented by an $M|G|\infty$ input process.

Thirdly, the $M|G|\infty$ model is extremely versatile in that, dependencies over a wide range of time scales can be exhibited simply by controlling the tail behaviour of σ [Prop. 3.1]: If $\Gamma(h)$ denotes the autocovariance of lag h for the stationary version of the $M|G|\infty$ process, then

$$\Gamma(h) = \lambda \mathbf{E}\left[\sigma\right] e^{-v_h}, \quad h = 0, 1, \ldots \tag{1.1}$$

where $v_h = -\ln \mathbf{P}\left[\widehat{\sigma} > h\right]$ and $\widehat{\sigma}$ is the forward recurrence time (2.5) associated with σ. This relation already indicates the tremendous amount of flexibility in modeling positive correlation structures. The degree of positive correlation exhibited by an $M|G|\infty$ input process can be further characterized by the sum of the autocovariances (1.1), or index of dispersion of counts (IDC). We show [Prop. 3.2] that

$$\text{IDC} \equiv \sum_{h=0}^{\infty} \Gamma(h) = \frac{\lambda}{2} \mathbf{E}\left[\sigma(\sigma+1)\right] \tag{1.2}$$

and the process is short–range dependent (i.e., IDC finite) if and only if $\mathbf{E}\left[\sigma^2\right]$ is finite.

Temporal correlations of $M|G|\infty$ input processes are known to affect queueing performance [17]. Insights into this phenomenon can be gained by analyzing the behaviour of a multiplexer fed by an $M|G|\infty$ input process. For simplicity, we model the multiplexer as a discrete–time single server system consisting of an infinite sized buffer and a server with a constant release rate c (cells/slot). The number of customers in the input buffer at time t is denoted by q_t. Our performance index is the steady–state buffer tail probability $\mathbf{P}\left[q_\infty > b\right]$, as this quantity is indicative of the buffer overflow probability in a corresponding finite buffer system with b positions.

Computing these tail probabilities, either analytically or numerically, represents a challenging problem in the absence of any underlying Markov property for $M|G|\infty$ inputs. Instead, we focus on the simpler task of determining the tail behaviour of the queue–length distribution in some asymptotic sense. More precisely, we seek results of the form

$$\lim_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}\left[q_\infty > b\right] = -\gamma^\star \tag{1.3}$$

for some positive constant $\gamma^\star$ and mapping $h : \mathbb{R}_+ \to \mathbb{R}_+$; these quantities are characterised by λ, G and c, and should be computable fairly easily. In Section 6 we carry out the calculations for several distributions, thin–tailed as well as heavy–tailed ones. Drastically different behavior emerge depending on whether $v_t = 0(t)$ or $v_t = o(t)$ [Thms. 5.1 and 5.2].

Limits such as (1.3) provide a fair idea of the tail of the queue–length distribution, and suggest approximations of the form

$$\mathbf{P}\left[q_\infty > b\right] \sim e^{-h(b)\gamma^\star} \quad (b \to \infty). \tag{1.4}$$

Of course, the use of the right handside of (1.4) to estimate $\mathbf{P}\left[q_\infty > b\right]$ may be fraught with difficulties [5]. Nonetheless, (1.3) already provides some qualitative insights into the queueing behavior at the multiplexer, and could in principle be used to produce guidelines for sizing up its buffers.

Our focus here is primarily on large deviations techniques in order to obtain (1.3). This approach has already been adopted by a number of authors [9, 14, 15]. Applying recent results by Duffield and O'Connell [9] we can compute $h(b)$ and $\gamma^\star$ under reasonably general conditions. Further, for a large class of distributions, we can select $h(b) = v_b$, and the asymptotics (1.3)–(1.4) then take the compact form

$$\mathbf{P}\left[q_\infty > b\right] \sim \mathbf{P}\left[\widehat{\sigma} > b\right]^{\gamma^\star} \quad (b \to \infty). \tag{1.5}$$

Hence, in many cases, including the Pareto, log–normal and Weibull service times, q_∞ and $\widehat{\sigma}$ (thus σ) belong to the same distributional class as characterized by tail behavior.

Sometimes, in lieu of (1.3), these large deviations techniques yield only weaker asymptotics in the form

$$\liminf_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}\left[q_\infty > b\right] \geq -\gamma^\star. \tag{1.6}$$

This situation typically occurs when σ is heavy–tailed, in which case large deviations excursions are only one of several causes for buffer exceedances [15]. Results such as (1.6) as still useful in that they provide *bounds* on decay rates.

Comparison of (1.5) with results from [20] and [22] points already to the complex and subtle impact of (long–range) dependencies on the tail probability $\mathbf{P}\left[q_\infty > b\right]$.

Indeed, in [20] the input stream to the multiplexer was modeled as a fractional Gaussian noise process exhibiting long–range dependence (in fact, self–similarity), and the buffer asymptotics displayed Weibull–like characteristics. On the other hand, by the results described above, an $M|G|\infty$ input process with a Weibull service time also yields Weibull–like buffer asymptotics although the input process is now short–range dependent. Hence, the *same* asymptotic buffer behavior can be induced by two vastly different input streams, one long–range dependent and the other short–range dependent! To make matters worse, if the pmf G were selected to be Pareto instead of Weibull, the input process would be long–range dependent, in fact asymptotically self–similar [22], but the buffer distribution would now exhibit Pareto–like asymptotics. To reiterate the main conclusion of [22], the value of the Hurst parameter as the sole indicator of long–range dependence (via asymptotic self–similarity) does not suffice for characterizing buffer asymptotics. Furthermore, buffer sizing cannot be determined adequately by appealing solely to the short versus long–range dependence characterization of the input model used, be it of the $M|G|\infty$ type or otherwise. Of course, long–range dependence (and its close cousin, self–similarity) are determined by second–order properties of the input process, while aymptotics of the form (1.3) invoke much finer probabilistic properties. The finiteness of $\mathbf{E}\left[\sigma^2\right]$ (needed in (1.2)) is obviously a poor marker for predicting the behavior of the sequence $\{v_t,\ t = 1, 2, \ldots\}$ (which drives (1.3)). To close, the diverse queueing behavior, demonstrated here and tied to the tail behavior of σ, not only confirms the versatility of $M|G|\infty$ inputs as network traffic models, but also points to the need for a very careful and cautious approach in modeling network traffic when time dependencies are either observed or suspected.

The paper is organized as follows: We introduce both the class of $M|GI|\infty$ inputs as well as the multiplexer model in Section 2 along with various preliminaries. We discuss the correlation structure of the $M|GI|\infty$ input process in Section 3. Section 4 develops general results on buffer asymptotics which are applied to our specific model in Section 5. Finally, Section 6 illustrates the asymptotic results for various selections of distribution function G.

A few words on the notation used in this paper: All rvs are defined on some probability triple $(\Omega, \mathcal{F}, \mathbf{P})$, with $\mathbf{E}$ denoting the corresponding expectation operator.

2 A multiplexer driven by $M|GI|\infty$ inputs

We model the multiplexer as a discrete–time single server queue with infinite buffer capacity which operates at a constant rate and in a first–come first–served manner: Let q_t denote the number of cells remaining in the buffer by the end of slot $[t-1,t)$, and let b_{t+1} denote the number of new cells which arrive at the start of time slot $[t,t+1)$. If the multiplexer output link can transmit c cells/slot, then the buffer content sequence $\{q_t,\ t=0,1,\ldots\}$ evolves according to Lindley recursion

$$q_0 = q; \quad q_{t+1} = [q_t + b_{t+1} - c]^+, \quad t = 0,1,\ldots \tag{2.1}$$

for some initial condition q.

Here, we account for time dependencies in the cell input stream by modelling the arrival process $\{b_t,\ t=0,1,\ldots\}$ as the busy server process of a discrete–time $M|G|\infty$ system. During time slot $[t,t+1)$, β_{t+1} new customers arrive into the system. Customer i, $i=1,\ldots,\beta_{t+1}$, is presented to its own server and begins service by the start of slot $[t+1,t+2)$; its service time has duration $\sigma_{t+1,i}$. Let b_t denote the number of busy servers, or equivalently of customers still present in the system, at the beginning of slot $[t,t+1)$, with b denoting the number of busy servers initially present in the system at $t=0$.

The $\mathbb{N}$–valued rvs b, $\{\beta_{t+1},\ t=0,1,\ldots\}$ and $\{\sigma_{t,i},\ t=0,1,\ldots;\ i=1,2,\ldots\}$ satisfy the following assumptions: (i) The rvs are mutually independent; (ii) The rvs $\{\beta_{t+1},\ t=0,1,\ldots\}$ are *i.i.d.* Poisson rvs with parameter $\lambda > 0$; (iii) The rvs $\{\sigma_{t,i},\ t=1,\ldots;\ i=1,2,\ldots\}$ are *i.i.d.* with common pmf G on $\{1,2,\ldots\}$. We denote by σ a generic $\mathbb{N}$–valued rv distributed according to the pmf G, and throughout we assume $\mathbf{E}[\sigma] < \infty$.

No additional assumptions are made on the rvs $\{\sigma_{0,i},\ i=1,2,\ldots\}$ which represent the (residual) service durations of the b customers present in the system at the beginning of the slot $[0,1)$. Various scenarios can thus be accommodated within this $M|G|\infty$ model: If the initial customers start their service at time $t=0$, then it is appropriate to assume that the rvs $\{\sigma_{0,i},\ i=1,2,\ldots\}$ are also i.i.d. rvs which are distributed according to the pmf G. On the other hand, if we take the viewpoint that the system has been in operation for some time, then these rvs $\{\sigma_{0,i},\ i=1,2,\ldots\}$ may be interpreted as the incomplete work (expressed in time slots) that the b "initial" customers require from their respective servers before their service is completed. In general, the statistics of the rvs $\{\sigma_{0,i},\ i=1,2,\ldots\}$ cannot be specified in any meaningful way, except for the situation that corresponds to the steady–state regime.

Many properties of $M|G|\infty$ input processes derive from the decomposition

$$b_t = b_t^{(0)} + b_t^{(a)}, \quad t = 0, 1, \ldots \tag{2.2}$$

where the rvs $b_t^{(0)}$ and $b_t^{(a)}$ describe the contributions to the number of customers in the system at the beginning of slot $[t, t+1)$ from those initially present (at $t = 0$) and from the new arrivals, respectively. It is plain that

$$b_t^{(0)} = \sum_{i=1}^{b} \mathbf{1}\left[\sigma_{0,i} > t\right], \quad t = 0, 1, \ldots \tag{2.3}$$

and that the rv $b_t^{(a)}$ can also be interpreted as the number of busy servers in the system at the beginning of slot $[t, t+1)$ given that the system was initially empty (i.e., $b = 0$).

In the next proposition we state conditions for the queueing system (2.1) to admit a steady–state regime when driven by the busy server process $\{b_t,\ t = 0, 1, \ldots\}$. Weak convergence is denoted by $\Longrightarrow$.

Proposition 2.1 *If $\lambda \mathbf{E}[\sigma] < c$, then there exists an $\mathbb{R}_+$–valued rv q_∞ such that $q_t \Longrightarrow_t q_\infty$ for any choice of the initial conditions q, b and $\{\sigma_{0,i},\ i = 1, 2, \ldots\}$. The system is then said to be stable.*

In general, the busy server process $\{b_t,\ t = 0, 1, \ldots\}$ is *not* a (strictly) stationary process, and Proposition 2.1 will not follow directly from the well–known stability result of Loynes [19] for Lindley recursions driven by stationary and ergodic sequences. The characterization of stability flows instead from an extension of Loynes' result to the case of driving sequences which *couple* with their stationary and ergodic versions [2]. To that end, we show [21] that the busy server process $\{b_t,\ t = 0, 1, \ldots\}$ indeed admits a stationary and ergodic version, thereafter denoted by $\{b_t^\star,\ t = 0, 1, \ldots\}$, and with which it couples for any choice of the initial conditions b and $\{\sigma_{0,i},\ i = 1, 2, \ldots\}$. This stationary version $\{b_t^\star,\ t = 0, 1, \ldots\}$ can represented through (2.3) with

$$b_t^{\star(0)} = \sum_{n=1}^{b} \mathbf{1}\left[\widehat{\sigma}_n > t\right], \quad t = 0, 1, \ldots \tag{2.4}$$

where (i) the rvs $\{\widehat{\sigma}_n,\ n = 1, 2, \ldots\}$ are independent of the rv b which is Poisson distributed with parameter $\lambda \mathbf{E}[\sigma]$, and (ii) the rvs $\{\widehat{\sigma}_n,\ n = 1, 2, \ldots\}$ are *i.i.d.* rvs distributed according to the forward recurrence time $\widehat{\sigma}$ associated with σ. This distribution is given by

$$\widehat{g}_r \equiv \mathbf{P}\left[\widehat{\sigma} = r\right] = \frac{\mathbf{P}\left[\sigma \geq r\right]}{\mathbf{E}\left[\sigma\right]}, \quad r = 1, 2, \ldots \tag{2.5}$$

Throughout, we assume at a minimum the conditions of Proposition 2.1, and write $r_{in} \equiv \lambda \mathbf{E}\left[\sigma\right]$ to stress the fact that $\lambda \mathbf{E}\left[\sigma\right]$ indeed represents the average input rate into the multiplexer. This will come in handy when comparing the effect on the multiplexer of several traffic streams distinguished by different distributions for σ with a given value for r_{in} such that $r_{in} < c$.

3 Correlation properties

We write

$$v_t \equiv -\ln \mathbf{P}\left[\widehat{\sigma} > t\right], \qquad t = 1, 2, \ldots \tag{3.1}$$

where the forward recurrence time $\widehat{\sigma}$ associated with the service time rv σ, is distributed according to (2.5). The following properties of $\{b_t^\star,\ t = 0, 1, \ldots\}$ are discussed in [6, 7, 21].

Proposition 3.1 *The stationary and ergodic version $\{b_t^\star,\ t = 0, 1, \ldots\}$ of the busy server process has the following properties:*

1. *For each $t = 0, 1, \ldots$, the rv $b_t^\star$ is a Poisson rv with parameter $\lambda \mathbf{E}\left[\sigma\right]$;*
2. *Its covariance structure is given by*

$$\begin{aligned} \Gamma(h) &\equiv \operatorname{cov}[b_t^\star, b_{t+h}^\star] \\ &= \lambda \mathbf{E}\left[(\sigma - h)^+\right] \\ &= \lambda \mathbf{E}\left[\sigma\right] e^{-v_h}, \quad t, h = 0, 1, 2, \ldots \end{aligned} \tag{3.2}$$

with the convention $v_0 \equiv 0$.

Proof. Fix $h = 1, 2, \ldots$: The first expression for $\Gamma(h)$ is well known [7]. Next, note that

$$\begin{aligned} \Gamma(h) &= \lambda \mathbf{E}\left[(\sigma - h)^+\right] \\ &= \lambda \sum_{r=0}^{\infty} \mathbf{P}\left[(\sigma - h)^+ > r\right] \end{aligned}$$

$$
\begin{aligned}
&= \lambda \sum_{r=0}^{\infty} \mathbf{P}\left[\sigma > h + r\right] \\
&= \lambda \sum_{r=h+1}^{\infty} \mathbf{P}\left[\sigma \geq r\right] \\
&= \lambda \mathbf{E}\left[\sigma\right] \sum_{r=h+1}^{\infty} \mathbf{P}\left[\widehat{\sigma} = r\right] \\
&= \lambda \mathbf{E}\left[\sigma\right] \mathbf{P}\left[\widehat{\sigma} > h\right]
\end{aligned}
$$

and (3.2) follows from (3.1). ■

The strength of the positive correlation exhibited by the sequence $\{b_t^\star,\ t = 0, 1, \ldots\}$ can be formalized as follows: We say that the sequence $\{b_t^\star,\ t = 0, 1, \ldots\}$ exhibits *short–range dependence* if

$$\sum_{h=0}^{\infty} \Gamma(h) < \infty. \tag{3.3}$$

Otherwise, the sequence $\{b_t^\star,\ t = 0, 1, \ldots\}$ is said to be *long-range dependent* [3, 4].

For $M|G|\infty$ processes this dependence can be characterized through the scaling $\{v_t,\ t = 1, 2, \ldots\}$, or alternatively through the finiteness of $\mathbf{E}\left[\sigma^2\right]$. First, from (3.2) we readily conclude that

Proposition 3.2 *Assume* $\lim_{t\to\infty} \frac{v_t}{\ln t} = K$. *If* $K \leq 1$ *(resp.* > 1*), then the stationary sequence* $\{b_t^\star,\ t = 0, 1, \ldots\}$ *is short-range (resp. long-range) dependent.*

Proposition 3.3 *We have the relation*

$$\sum_{h=0}^{\infty} \Gamma(h) = \lambda \mathbf{E}\left[\sigma\right] \mathbf{E}\left[\widehat{\sigma}\right] = \frac{\lambda}{2} \mathbf{E}\left[\sigma(\sigma + 1)\right], \tag{3.4}$$

so that the stationary sequence $\{b_t^\star,\ t = 0, 1, \ldots\}$ *is short-range (resp. long-range) dependent if and only if* $\mathbf{E}\left[\sigma^2\right]$ *is finite (resp. infinite).*

Proof. From (3.2) we see that

$$
\begin{aligned}
\sum_{h=0}^{\infty} \Gamma(h) &= \lambda \mathbf{E}\left[\sigma\right] \sum_{h=0}^{\infty} \mathbf{P}\left[\widehat{\sigma} > h\right] \\
&= \lambda \mathbf{E}\left[\sigma\right] \mathbf{E}\left[\widehat{\sigma}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \lambda \mathbf{E}[\sigma] \sum_{r=1}^{\infty} r \mathbf{P}[\widehat{\sigma} = r] \\
&= \lambda \mathbf{E}[\sigma] (\mathbf{E}[\sigma])^{-1} \sum_{r=1}^{\infty} r \mathbf{P}[\sigma \geq r] \\
&= \lambda \sum_{r=1}^{\infty} r \sum_{t=r}^{\infty} \mathbf{P}[\sigma = t] \\
&= \lambda \sum_{t=1}^{\infty} \mathbf{P}[\sigma = t] \left(\sum_{r=1}^{t} r \right) \\
&= \frac{\lambda}{2} \sum_{t=1}^{\infty} t(t+1) \mathbf{P}[\sigma = t]
\end{aligned}
$$

and the conclusion (3.4) is now immediate. ■

4 General Results on Buffer Asymptotics

Several authors [9, 14, 15] have derived asymptotics such as (1.3) by means of large deviations estimates associated with the sequence $\{S_t - ct,\ t = 0, 1, \ldots\}$, where

$$S_0 = 0; \quad S_t = b_1^\star + \ldots + b_t^\star, \quad t = 1, 2, \ldots \tag{4.1}$$

These results have been obtained in varying degrees of generality, and are summarized below as they apply to the present context.

To fix the terminology, consider a monotone increasing $\mathbb{R}$–valued sequence $\{v_t,\ t = 0, 1, \ldots\}$ such that $\lim_{t\to\infty} v_t = \infty$. A sequence of $\mathbb{R}$–valued rvs $\{x_t,\ t = 0, 1, \ldots\}$ is said to satisfy the *Large Deviations Principle under scaling* v_t if there exists a lower–semicontinuous function $I : \mathbb{R} \to [0, \infty]$ such that for every open set G,

$$-\inf_{x \in G} I(x) \leq \liminf_{t\to\infty} \frac{1}{v_t} \ln \mathbf{P}[x_t \in G] \tag{4.2}$$

and for every closed set F,

$$\limsup_{t\to\infty} \frac{1}{v_t} \ln \mathbf{P}[x_t \in F] \leq -\inf_{x \in F} I(x). \tag{4.3}$$

The rate function I is said to be good if for each $r > 0$, the level set $\{x \in \mathbb{R} : I(x) \leq r\}$ is a compact subset of $\mathbb{R}$.

In many situations of interest, the rate function can be expressed as the Legendre–Fenchel transform $\Lambda^\star$ of another mapping $\Lambda : \mathbb{R} \to (-\infty, \infty]$, namely

$$\Lambda^\star(z) \equiv \sup_{\theta \in \mathbb{R}} \{\theta z - \Lambda(\theta)\}, \quad z \in \mathbb{R}. \tag{4.4}$$

The reader is referred to the monograph [8] for additional information on the subject matter of Large Deviations.

Consider two monotone increasing $\mathbb{R}_+$–valued sequences $\{v_t,\ t = 0, 1, \ldots\}$ and $\{a_t,\ t = 0, 1, \ldots\}$ increasing at infinity, i.e., $\lim_{t\to\infty} v_t = \lim_{t\to\infty} a_t = \infty$, For each $t = 1, 2, \ldots$, we define

$$\Lambda_t(\theta) \equiv \frac{1}{v_t} \ln \mathbf{E}\left[\exp\left(\theta v_t \frac{S_t - ct}{a_t}\right)\right], \quad \theta \in \mathbb{R}. \tag{4.5}$$

If we can show the existence of functions $g, h : \mathbb{R}_+ \to \mathbb{R}_+$ such that h is monotone increasing with $\lim_{b\to\infty} h(b) = \infty$ and the limit

$$\lim_{b\to\infty} \frac{v_{[b/y]}}{h(b)} = g(y), \quad y > 0 \tag{4.6}$$

holds, then the following theorem obtained by Duffield and O'Connell in [9] applies.

Proposition 4.1 *Assume the arrival sequence $\{b_{t+1},\ t = 0, 1, \ldots\}$ to be stationary and ergodic, and to satisfy the following conditions:*

1. *For each θ in $\mathbb{R}$, the limit $\Lambda(\theta) \equiv \lim_{t\to\infty} \Lambda_t(\theta)$ exists (possibly as an extended real number);*

2. *The process $\{a_t^{-1}(S_t - ct),\ t = 1, 2, \ldots\}$ satisfies the Large Deviations Principle with good rate function $\Lambda^\star$ under scaling v_t.*

Then, for each $y > 0$ we have

$$\liminf_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}[q_\infty > b] \geq -g(y) \inf_{x>y} \Lambda^\star(x). \tag{4.7}$$

If the (convex) rate function $\Lambda^\star$ is continuous on $[0, \infty)$, then Proposition 4.1 immediately implies the lower bound

$$\liminf_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}[q_\infty > b] \geq -\gamma^\star \tag{4.8}$$

with $\gamma^\star$ given by

$$\gamma^\star \equiv \inf_{y>0} g(y) \Lambda^\star(y). \tag{4.9}$$

In [9], under additional conditions to the ones of Proposition 4.1, a companion upper bound to (4.7)–(4.8) is derived. Proposition 4.2 below incorporates these conditions with a few adjustments. A different approach to calculating the upper bound is discussed in [24], and yields identical bounds.

Proposition 4.2 *Assume the arrival sequence $\{b_{t+1},\ t = 0, 1, \ldots\}$ satisfies the specifications of Proposition 4.1 as well as the following conditions:*

1. $\inf_{x>0} g(x)\Lambda^\star(x) < \infty$*;*

2. *For every $\gamma > 0$, there exists $y = y(\gamma) > 0$ such that*

i.

$$\limsup_{b\to\infty} \frac{1}{h(b)} \ln \left(\sum_{k=a^{-1}(\frac{b}{y})}^{\infty} e^{-\gamma v_k} \right) \leq - \inf_{x>0} g(x)\Lambda^\star(x);$$

ii.

$$\limsup_{b\to\infty} \frac{1}{h(b)} \ln a^{-1}(\frac{b}{y}) = 0;$$

iii.

$$\lim_{t\to\infty} \inf_{x>y} \frac{\Lambda^\star(x) v_t}{h(x a_t)} = \inf_{x>y} g(x)\Lambda^\star(x).$$

Then, we have

$$\limsup_{b\to\infty} \frac{1}{h(b)} \ln \mathbf{P}\left[q_\infty > b\right] \leq - \inf_{y>0} g(y)\Lambda^\star(y). \tag{4.10}$$

When the conditions of both Propositions 4.1 and 4.2 are satisfied, the asymptotics distribution take the form (1.3) (with $\gamma^\star$ usually given by (4.9)).

When the pmf G is heavy–tailed, the conditions of Proposition 4.1 hold whereas those of Proposition 4.2 do *not*, leaving us without an upper bound. Going back to the heuristics given in [15], we attribute this to the fact that now buffer exceedances cannot be explained entirely by large deviations excursions in the arrival stream, as there is a need to take into consideration the effect of a single customer with a large workload. Hence, any argument based on large deviations techniques *alone* is bound to fall short. However, we conjecture that (1.3) still holds with scaling h as specified through (4.6) but of course with a different value for $\gamma^\star$. Work on this issue is in progress.

5 Evaluation of $\gamma^\star$ and $h(b)$

The key step in applying the results of the previous section consists in finding two monotone increasing $\mathbb{R}_+$–valued sequences $\{v_t,\ t = 0, 1, \ldots\}$ and $\{a_t,\ t = 0, 1, \ldots\}$

increasing at infinity, such that the limit

$$\Lambda(\theta) \equiv \lim_{t\to\infty} \Lambda_t(\theta), \quad \theta \in \mathbb{R} \tag{5.1}$$

exists (possibly as an extended real number). It is helpful to view a_t as the scaling representative of a law of large numbers and v_t as a large deviations type scaling.

In the context of $M|G|\infty$ processes, it is natural to begin with the selection $a_t = t$. In the paper [23], we show that the appropriate large deviations scaling $\{v_t,\ t = 1, 2, \dots\}$ is given by (3.1). To state the results more conveniently, we set

$$\Lambda_{b,t}(\theta) \equiv \frac{1}{v_t} \ln \mathbf{E}\left[\exp(\frac{v_t}{t}\theta S_t)\right], \quad \theta \in \mathbb{R} \tag{5.2}$$

for each $t = 1, 2, \dots$. Obviously, if the limit

$$\Lambda_b(\theta) \equiv \lim_{t\to\infty} \Lambda_{b,t}(\theta), \quad \theta \in \mathbb{R} \tag{5.3}$$

exists, so does (5.1) with

$$\Lambda(\theta) = \Lambda_b(\theta) - c\theta, \quad \theta \in \mathbb{R} \tag{5.4}$$

and it suffices to concentrate on finding (5.3). The main facts along these lines are developed in the next two theorems; proofs are available in [21, 23].

Theorem 5.1 *Assume $v_t = O(t)$ with $\lim_{t\to\infty} v_t/t = R > 0$. Then, for each θ in $\mathbb{R}$, the limit $\Lambda_b(\theta) \equiv \lim_{t\to\infty} \Lambda_{b,t}(\theta)$ exists and is given by*

$$\Lambda_b(\theta) = \begin{cases} \lambda \mathbf{E}\left[\sigma\right] \left(\frac{e^{R\theta}-1}{R}\right) \Sigma(\theta) & \text{\textit{if} } \theta < 1 \\ \infty & \text{\textit{if} } \theta > 1 \end{cases} \tag{5.5}$$

where

$$\Sigma(\theta) = 1 + \left(1 - e^{-R\theta}\right)\left(\sum_{r=1}^{\infty} \exp\left(r(\theta R - \frac{v_r}{r})\right)\right), \quad \theta \in \mathbb{R}. \tag{5.6}$$

Moreover, $\Sigma(\theta)$ is finite for $\theta < 1$.

We say that the sequence $\{v_t/t,\ t = 1, 2, \dots\}$ is monotone decreasing in the limit if there exists a finite integer T such that the tail $\{v_t/t,\ t = T+1, T+2, \dots\}$ is monotone decreasing.

Theorem 5.2 *Assume $v_t = o(t)$ with $\{v_t/t,\ t = 1, 2, \dots\}$ monotone decreasing in the limit. Assume further that there exists a mapping $\Gamma : \mathbb{N} \to \mathbb{N}$ such that (i)*

$\Gamma(t) < t$ *for all* $t = 1, 2, \ldots$, *(ii)* $\lim_{t\to\infty} v_t \frac{\Gamma(t)}{t} = \infty$ *and (iii)* $\lim_{t\to\infty} \frac{v_t}{t} \frac{\Gamma(t)}{v_{\Gamma(t)}} = 0$. *Then, for each* θ *in* $\mathbb{R}$, *the limit* $\Lambda_b(\theta) \equiv \lim_{t\to\infty} \Lambda_{b,t}(\theta)$ *exists and is given by*

$$\Lambda_b(\theta) = \begin{cases} \lambda \mathbf{E}[\sigma]\,\theta & \text{if } \theta < 1 \\ \infty & \text{if } \theta > 1. \end{cases} \tag{5.7}$$

Neither of the above–mentioned theorems deals with the case $\theta = 1$. However, it can be shown [21] that $\lim_{t\to\infty} \Lambda_{b,t}(1)$ does exist for the examples discussed here. In fact, a little thought indicates that the *existence* of this limit suffices for our purpose, in that its *specific* value is not of any consequence in evaluating $\gamma^\star$ as given by (4.9). Indeed, in that case, under the assumptions of either Theorem 5.1 or 5.2, we see from (4.4) and (5.4) that

$$\begin{aligned} \Lambda^\star(z) &= \sup_{\theta\in\mathbb{R}} \{\theta z - (\Lambda_b(\theta) - c\theta)\} \\ &= \sup_{\theta\le 1} \{(c+z)\theta - \Lambda_b(\theta)\} \\ &= \sup_{\theta< 1} \{(c+z)\theta - \Lambda_b(\theta)\}, \quad z \in \mathbb{R} \end{aligned} \tag{5.8}$$

where in the last step we have used the fact $\lim_{\theta\uparrow 1} \Lambda_b(\theta) \le \Lambda_b(1)$. Moreover, in the case $v_t = o(t)$, further computations are possible under Theorem 5.2: From (5.8) and the stability condition $\lambda \mathbf{E}[\sigma] = r_{in} < c$, we get

$$\begin{aligned} \Lambda^\star(z) &= \sup_{\theta<1} (z - (\lambda \mathbf{E}[\sigma] - c))\,\theta \\ &= z + c - r_{in}, \quad z > 0, \end{aligned} \tag{5.9}$$

whence

$$\gamma^\star = \inf_{y>0} g(y)(y + c - r_{in}). \tag{5.10}$$

As becomes apparent through the examples discussed in [24], the term v_t often takes on a form which is computationally inconvenient when checking the various technical conditions. Fortunately, only the *asymptotic* behavior of v_t matters. More precisely, consider another $\mathbb{R}_+$–valued sequence $\{w_t,\ t = 1, 2, \ldots\}$ which is asymptotically equivalent to $\{v_t,\ t = 1, 2, \ldots\}$ in the sense that

$$\lim_{t\to\infty} \frac{w_t}{v_t} = 1. \tag{5.11}$$

In reference to (4.6), we now seek mappings $h, g : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\lim_{b\to\infty} \frac{w_{\lfloor b/y \rfloor}}{h(b)} = g(y), \quad y > 0. \tag{5.12}$$

It is then easy to check from (5.11) that

$$\lim_{b\to\infty} \frac{v_{[b/y]}}{h(b)} = \lim_{b\to\infty} \frac{w_{[b/y]}}{h(b)} \lim_{b\to\infty} \frac{v_{[b/y]}}{w_{[b/y]}} = g(y), \quad y > 0 \tag{5.13}$$

and these mappings also satisfy (4.6)! A similar approach applies when checking conditions (i)–(iii) in Theorem 5.2. Here, suppose that we have found a mapping $\Gamma : I\!N \to I\!N$ such that (i) $\Gamma(t) < t$ for all $t = 1, 2, \ldots$, (ii) $\lim_{t\to\infty} w_t \frac{\Gamma(t)}{t} = \infty$ and (iii) $\lim_{t\to\infty} \frac{w_t}{t} \frac{\Gamma(t)}{w_{\Gamma(t)}} = 0$ – in other words, conditions (i)–(iii) with respect to w_t. Again, the asymptotic equivalence (5.11) implies that Γ satisfies conditions (i)–(iii) (with respect to v_t) in Theorem 5.2.

In short, in applying the results of Sections 4 and 5, we see that the appropriate conditions can all be checked by replacing v_t by any other scaling w_t which is asymptotically equivalent to it in the sense of (5.11), and hopefully more tractable analytically. We shall refer to any such scaling as an auxiliary scaling, and will use it to check various conditions.

If the function $u : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying and monotone, then

$$\lim_{b\to\infty} \frac{u([b/y])}{u(b)} = g(y), \quad y > 0; \tag{5.14}$$

in fact, it is well known [11] that

$$g(y) = y^\rho, \quad y > 0; \ -\infty \le \rho \le \infty. \tag{5.15}$$

This suggests a natural choice for h as follows: Denote by $w : \mathbb{R}_+ \to \mathbb{R}_+$ the piecewise–continuous inperpolation of the auxiliary scaling sequence $\{w_t, \ t = 0, 1, \ldots\}$ (with $w_0 = 0$). If w is regularly varying, then we can select $h(b) = w(b)$ for all $b > 0$, in which case the asymptotics (1.4) takes the form

$$\mathbf{P}\left[q_\infty > b\right] \sim \mathbf{P}\left[\widehat{\sigma} > b\right]^{\gamma^\star} \quad (b \to \infty). \tag{5.16}$$

A closer look at (5.16) indicates that for a number of distributions the tail behaviours of the service time σ and of the queue length q_∞ are of the same type.

6 Examples

We illustrate these asymptotic results for various choices of the pmf G; in each case we identify $h(b)$ and compute $\gamma^\star$ in closed form (whenever possible). When a simple expression is not provided for $\gamma^\star$, it should be understood that its numerical value can obtained by solving an easy one–dimensional optimization problem. The results are presented in order of increasing weight on the tail of G, or equivalently, in order of increasing time dependence. We begin with the Rayleigh distribution which exhibits the least degree of time dependence and proceed to discuss the Gamma, geometric, Weibull and log–normal cases, all of which are short range dependent. Finally, we discuss the Pareto distribution which exhibits long–range dependence (in fact asymptotic self–similarity). Details of the calculations are available in [24].

A rv X is described as having a Rayleigh distribution with parameter $\alpha > 0$ if

$$\mathbf{P}[X \leq x] = 1 - e^{-\frac{x^2}{2\alpha^2}}, \quad x \geq 0. \tag{6.1}$$

The pmf $G = \{g_r,\ r = 1, 2, \ldots\}$ of the rv σ is said to be an (integer–valued) Rayleigh distribution with parameter $\alpha > 0$, if $\sigma =_{st} \lceil X \rceil$, whence

$$g_r = e^{-\frac{(r-1)^2}{2\alpha^2}} - e^{-\frac{r^2}{2\alpha^2}}, \quad r = 1, 2, \ldots \tag{6.2}$$

Proposition 6.1 *If G is a discrete Rayleigh distribution with parameter α where $\alpha > 0$, then*

$$\lim_{b \to \infty} \frac{1}{b^2} \ln \mathbf{P}[q_\infty > b] = -\gamma^\star_{\text{Rayleigh}} \tag{6.3}$$

for some finite constant $\gamma^\star_{\text{Rayleigh}}$.

A rv X has a Gamma distribution with parameters $c > 0$ and $b > 0$ if

$$\mathbf{P}[X \leq x] = 1 - \frac{Q(b, cx)}{\Gamma(b)}, \quad x \geq 0 \tag{6.4}$$

where

$$Q(\eta, x) \equiv \int_x^\infty e^{-t} t^{\eta - 1} dt, \quad \eta \geq 0,\ x > 0$$

is the incomplete Γ–function and $\Gamma(\eta) \equiv Q(\eta, 0)$.

The pmf $G = \{g_r,\ r = 1, 2, \ldots\}$ of the rv σ is said to be an (integer–valued) Gamma distribution with parameters $c > 0$ and $b > 0$ if $\sigma =_{st} \lceil X \rceil$, which yields

$$g_r = \frac{1}{\Gamma(b)} \left(Q(b, c(r-1)) - Q(b, cr) \right), \quad r = 1, 2, \ldots \tag{6.5}$$

Proposition 6.2 *If G is a discrete Gamma distribution with parameters $c > 0$ and $b > 0$, then*

$$\lim_{b \to \infty} \frac{1}{b} \ln \mathbf{P}[q_\infty > b] = -\gamma^\star_{\text{Gamma}} \tag{6.6}$$

where $\gamma^\star_{\text{Gamma}}$ is a finite constant.

The geometric pmf $G = \{g_r,\ r = 1, 2, \ldots\}$ with parameter $0 < q < 1$, is given by

$$g_r \equiv \mathbf{P}[\sigma = r] = (1-q)q^{r-1}, \quad r = 1, 2, \ldots \tag{6.7}$$

Proposition 6.3 *If G is a geometric pmf of parameter q, with $0 < q < 1$, given by (6.7), then*

$$\lim_{b\to\infty} \frac{1}{b} \ln \mathbf{P}[q_\infty > b] = -\gamma^\star_{\text{Geometric}} \tag{6.8}$$

where $\gamma^\star_{\text{Geometric}}$ is a finite constant.

A rv X is said to be a Weibull rv with parameters a and β, $a > 0$ and $0 < \beta < 1$, if

$$\mathbf{P}[X \le x] = 1 - e^{-ax^\beta}, \quad x \ge 0 \tag{6.9}$$

The pmf $G = \{g_r,\ r = 1, 2, \ldots\}$ of the rv σ is said to be an (integer–valued) Weibull distribution with parameters a and β if $\sigma =_{st} \lceil X \rceil$, in which case we have

$$g_r = e^{-a(r-1)^\beta} - e^{-ar^\beta}, \quad r = 1, 2, \ldots \tag{6.10}$$

Proposition 6.4 *If G is a discrete Weibull distribution with parameters a and β, $a > 0$ and $0 < \beta < 1$, then*

$$\lim_{b\to\infty} \frac{1}{b^\beta} \ln \mathbf{P}[q_\infty > b] = -\gamma^\star_{\text{Weibull}} \tag{6.11}$$

where

$$\gamma^\star_{\text{Weibull}} = \frac{a}{\beta}\left(\frac{\beta(c - r_{in})}{1-\beta}\right)^{1-\beta}. \tag{6.12}$$

A rv X is said to be a log–normal rv if $X =_{st} \exp(Y)$ where Y is a Gaussian rv with mean μ and variance δ^2. The pmf $G = \{g_r,\ r = 1, 2, \ldots\}$ of the rv σ is said to be an (integer–valued) log–normal distribution with parameters μ and δ if $\sigma =_{st} \lceil X \rceil$. It is easy to check that

$$\begin{aligned} g_r &= \mathbf{P}[r - 1 < X \le r] \\ &= \Phi\left(\frac{1}{\delta} \ln\left(\frac{r}{m}\right)\right) - \Phi\left(\frac{1}{\delta} \ln\left(\frac{r-1}{m}\right)\right), \quad r = 1, 2, \ldots \end{aligned} \tag{6.13}$$

where $m \equiv e^\mu$, and Φ is the cumulative distribution function of a Gaussian rv with zero mean and unit variance.

Proposition 6.5 *If G is a discrete log-normal distribution with parameters μ and δ as described above, then*

$$\liminf_{b\to\infty} \frac{1}{(\ln b)^2} \ln \mathbf{P}\left[q_\infty > b\right] \geq -\gamma^\star_{\text{Lognormal}} \tag{6.14}$$

where

$$\gamma^\star_{\text{Lognormal}} = \frac{c - r_{in}}{2\delta^2}. \tag{6.15}$$

In all the cases considered so far, $\mathbf{E}\left[\sigma^2\right]$ is finite, and the process $\{b^\star_t, t = 0, 1, \ldots\}$ is therefore short–range dependent by virtue of Proposition 3.3. However, in the log-normal case, we have only obtained the lower bound. Going beyond the technical conditions, we explain this departure from previous short–range dependent cases by the following argument: Although the log–normal distribution has finite second moment, thereby insuring short–range dependence (though barely as it turns out), its tail has become too heavy ($v_t = o(t)$) to neglect the effect of a single cutomer with a large workload.

Finally, the pmf $G = \{g_r,\ r = 1, 2, \ldots\}$ is said to be a Pareto distribution with parameter α, $1 < \alpha < 2$, if

$$\lim_{r\to\infty} \frac{\mathbf{P}\left[\sigma > r\right]}{r^{-\alpha}} = 1. \tag{6.16}$$

For the sake of concreteness we use

$$g_r = \mathbf{P}\left[\sigma = r\right] = r^{-\alpha} - (r+1)^{-\alpha}, \quad r = 1, 2, \ldots \tag{6.17}$$

so that $\mathbf{E}\left[\sigma\right] < \infty$ while $\mathbf{E}\left[\sigma^2\right] = \infty$, and we conclude to the long–range dependence of the process $\{b^\star_t, t = 0, 1, \ldots\}$.

Proposition 6.6 *If G is a discrete Pareto distribution with parameter α, $1 < \alpha < 2$, then*

$$\liminf_{b\to\infty} \frac{1}{\ln b} \ln \mathbf{P}\left[q_\infty > b\right] \geq -\gamma^\star_{\text{Pareto}} \tag{6.18}$$

where

$$\gamma^\star_{\text{Pareto}} = (\alpha - 1)(c - r_{in}). \tag{6.19}$$

As a final word we recall our earlier comments at the conclusion of Section 4; when the service time is heavy–tailed as illustrated by the log–normal or Pareto cases, the conditions of Proposition 4.2 are not all satisfied. The investigation of the buffer asymptotics for such processes will require that we look beyond the large deviation techniques used thus far.

References

[1] R.G. Addie, M. Zukerman and T. Neame, "Fractal traffic: Measurements, modeling and performance evaluation," in *Proceedings of Infocom '95*, Boston (MA), April 1995, pp. 985–992.

[2] F. Baccelli and P. Bremaud, *Elements of Queueing Theory: Palm–Martingale Calculus and Stochastic Recurrences*, Applications of Mathematics **26**, Springer–Verlag, Berlin Heidelberger, 1994.

[3] J. Beran, *Statistics for Long-Memory Processes*, Chapman and Hall, New York (NY), 1994.

[4] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger "Long-range dependence in variable bit–rate video traffic," *IEEE Transactions on Communications* **COM–43** (1995), pp. 1566–1579.

[5] G.L. Choudhury, D.M. Lucantoni and W. Whitt, "Sqeezing the most out of ATM," *IEEE Transactions on Communications* **COM–44** (1996), pp. 203–217.

[6] D. R. Cox, "Long–Range Dependence: A Review," *Statistics: An Appraisal*, H. A. David and H. T. David, Eds., The Iowa State University Press, Ames (IA), 1984, pp. 55–74.

[7] D. R. Cox and V. Isham, *Point Processes*, Chapman and Hall, New York (NY), 1980.

[8] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Jones and Bartlett, Boston (MA), 1993.

[9] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with applications," *Mathematical Proceedings of the Cambridge Philosophical Society* **118** (1995), pp. 363–374.

[10] A. Erramilli, O. Narayan and W. Willinger, "Experimental queuing analysis with long–range dependent packet traffic," *IEEE/ACM Transactions on Networking* **4** (1996), pp. 209–223.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II*, J. Wiley & Sons, New York (NY), 1966.

[12] H. J. Fowler and W. E. Leland, "Local area network traffic characteristics, with implications for broadband network congestion management," *IEEE Journal on Selected Areas in Communications* **JSAC–9** (1991), pp. 1139–1149.

[13] M. Garrett and W. Willinger, "Analysis, modeling and generation of self–similar VBR video traffic," *Proceedings of SIGCOMM '94*, September 1994, pp. 269–280.

[14] P.W. Glynn and W. Whitt, "Logarithmic asymptotics for steady–state tail probabilities in a single–server queue," *Journal of Applied Probability* **31** (1994), pp. 131–159.

[15] G. Kesidis, J. Walrand and C.S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Transactions on Networking* **1** (1993), pp. 424–428.

[16] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self–similar nature of ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking* **2** (1994), pp. 1–15.

[17] M. Livny, B. Melamed and A.K. Tsiolis, "The impact of autocorrelation on queueing systems," *Management Science* **39** (1993), pp. 322–339.

[18] N. Likhanov, B. Tsybakov and N.D. Georganas, "Analysis of an ATM buffer with self–similar ("fractal") input traffic," in *Proceedings of Infocom '95*, Boston (MA), April 1995, pp. 985–992.

[19] R.M. Loynes, "The stability of a queue with non–independent inter–arrival and service times," *Proceedings of the Cambridge Philosophical Society* **58** (1962), pp. 497–520.

[20] I. Norros, "A storage model with self–similar input," *Queueing Systems – Theory & Applications* **16** (1994), pp. 387-396.

[21] M. Parulekar, *Buffer Engineering for Self–Similar Traffic*, Ph.D. Thesis, Electrical Engineering Department, University of Maryland, College Park (MD). Expected December 1996.

[22] M. Parulekar and A.M. Makowski, "Tail probabilities for a multiplexer with self–similar traffic," Infocom'96, April 1996, San Francisco (CA).

[23] M. Parulekar and A.M. Makowski, "Tail probabilities for $M|G|\infty$ input processes (I): Preliminary asymptotics," *Queueing Systems – Theory & Applications*, submitted (1996).

[24] M. Parulekar and A.M. Makowski, "Tail probabilities for $M|G|\infty$ input processes (II): Asymptotics" in preparation.

[25] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Transactions on Networking* **3** (1993), pp. 226–244.

[26] W. Willinger, M. S. Taqqu, W. E. Leland and D. V. Wilson, "Self–similarity in high–speed packet traffic: Analysis and modeling of ethernet traffic measurements," *Statistical Science*, to appear.